

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICATION FOR LETTERS PATENT

**Load Simulation Tool For Server
Resource Capacity Planning**

Inventor(s):
Matt Odhner
Giedrius Zizys

ATTORNEY'S DOCKET NO. MS1-517US

0057448-05390

1 **TECHNICAL FIELD**

2 This invention relates to server systems, and more particularly to systems
3 and methods for server resource capacity planning in server systems.
4

5 **BACKGROUND**

6 Capacity planning is forward-looking resource management that allows a
7 computer system administrator to plan for expected changes of system resource
8 utilization and make changes to the system to adequately handle such changes.
9 Server performance and capacity planning is a top concern of computer
10 administrators and business managers. If a lack of proactive and continuous
11 capacity planning procedure leads to unexpected unavailability and performance
12 problems, the downtime that results can be financially devastating to a company
13 that depends heavily on server performance, such as an Internet-based merchant.

14 The importance of superior capacity planning is heightened by the
15 continuous growth in server-dependent companies and potential customers for
16 such companies. Even a solid company that has millions of customers can quickly
17 decline in popularity if it does not increase its resources to handle a constant
18 increase in customers. Excessive downtime of such a company can cause
19 customers to take their business elsewhere.

20 Capacity planning requires both scientific and intuitive knowledge of a
21 server system. It requires in-depth knowledge of the resource being provided and
22 an adequate understanding of future server traffic. The difficulty of the problem
23 has increased by technology in which multiple servers, or server clusters, are
24 employed to handle a network or an Internet website.
25

1 Current capacity planning methods do not adequately estimate a number of
2 servers having certain resources that a system will need to handle expected loads
3 (requests per second). Therefore, a capacity planning method and system is
4 needed in which a user can provide an expected load that the system needs to
5 handle and receive information on how to increase servers and/or resources to
6 adequately handle that load.

7 8 SUMMARY

9 A method and system for providing capacity planning of server resources is
10 described herein. The methods and systems contemplate using measured data,
11 extrapolation, and a load simulation tool to provide capacity planning results that
12 are more accurate than current schemes. The load simulation tool and its
13 implementation are also described. Server resources for which utilization is
14 calculated are processor utilization, communication bandwidth utilization,
15 memory utilization, and general server utilization.

16 Utilization is expressed in terms of actual use of the resource in relation to
17 the total amount of resource available for use. For example, processor utilization
18 is expressed as a percentage of procession power utilized for a given load in
19 relation to the total processing power available. Communication bandwidth
20 utilization is expressed as a percentage of an average server throughput per bytes
21 per second in relation to the total communication bandwidth available. Memory
22 utilization is expressed as a percentage of memory required per request times the
23 length of a request queue in relation to the total memory available. General server
24 utilization is expressed as a ratio between a current service rate (number of
25 requests per second served) and the maximum possible service rate (maximum

1 number of requests the server is capable of serving). This is less specific than
2 showing the processor, bandwidth, and memory utilization, but it is useful for
3 viewing resource constraints that do not fall under the other three categories.

4 The calculations that are used to derive utilization percentages of server
5 resources require that the maximum load that can be handled by the server cluster
6 (maximum requests / second) be determined. Other methods to estimate this
7 maximum load are described in a related patent application entitled, "Capacity
8 Planning For Server Resources," by Odhner and Zizys, U.S. Patent Application
9 No. 08/_____, filed _____. It is noted that the inventors
10 of the referenced patent application are the same of those of the present
11 application, and that Microsoft Corp. is the assignee of both inventions.

12 The implementation described herein derives the maximum load of a server
13 cluster by collecting actual server parameter values during operation of the server
14 system. This is accomplished through the use of a filter, such as an Internet Server
15 Application Program Interface (ISAPI) filter, that collects actual server traffic
16 information as data is transmitted to and from the server cluster. In addition, a
17 monitor on each server in the server cluster collects other server parameter values
18 that are used in subsequent calculations.

19 After the filter and the monitors have collected the required data, a system
20 user selects a client computer from which to run a load simulation tool. The load
21 simulation tool, in effect, replays the data that has been collected from the server
22 cluster, such as the actual requests made to the server, the time intervals at which
23 requests were made, etc. The load simulation tool is then used to increase the load
24 on the system until a maximum service rate that the system can support is found.
25

There are several ways to calibrate the server load to find the maximum service rate. The number of users from the actual recorded data can be multiplied to simulate a greater number of users, which will increase the load on the system. Another way is to decrease the amount of time between requests, as recorded by the system, which will increase the load on the system. As the load increases, a service rate is monitored. When a further increase in the load does not increase the service rate, the load on the system at that point is considered to be the maximum service rate that can be delivered by the server.

It is noted that the user can create a script manually, instead of replaying the recorded data to calibrate the maximum load, but this will not provide a similarly accurate outcome, since the user in that situation, is required to estimate certain server usage parameters.

After the system is calibrated to find the maximum load that can be handled by the system, the maximum load value is used in subsequent calculations to determine server resource utilization estimates for any number of hypothetical situations. For instance, a user can enter information regarding a particular load that the user wants the current system to handle. The described implementation provides that user with estimates as to the utilization that the specified load will cause for the processor, the memory, the communications bandwidth, and the server in general. Also, the user may want to see how adding or removing a server from a current system will affect the utilization of these server resources. This situation can be adequately determined using the implementation described herein.

Finally, after the user runs the load simulation tool to calibrate the system as to the maximum load and make determinations regarding utilization of server resources, the system provides a plan that recommends any changes in

configuration, if any, that should be made to the system to optimize system performance. These recommendations are stored for each test result, thereby enabling the user to run several tests, and contrast and compare results and recommendations for different situations that the user may expect in the future. The user is thus enabled to adequately plan for future situations.

BRIEF DESCRIPTION OF THE DRAWINGS

A more complete understanding of the various methods and arrangements of the present invention may be had by reference to the following detailed description when taken in conjunction with the accompanying drawings, wherein:

Fig. 1 is an illustration of a prior art server-client system having a server cluster that supports a website on the Internet.

Fig. 2 is a high-level block diagram of a server cluster having a stress simulation tool for capacity planning.

Fig. 3 is a screen shot of a capacity planning worksheet utilized in a capacity planning process using a stress simulation tool.

Fig. 4 is a graph of load vs. processor utilization for a calibrated method of capacity planning.

DETAILED DESCRIPTION

Fig. 1 shows a typical Internet-based server-client system 100. The system 100 includes several clients 104a, 104b, 104c, 104d connected to the Internet 102. A website 106 runs on a server cluster 108 comprised of three servers 110a, 110b, 110c. Although the server-client system 100 is shown operating within an Internet website context, it is noted that the server-client system may operate in any server-

client network context, such as a local area network (LAN) or a wide area network (WAN).

Fig. 2 depicts a server cluster 200 in accordance with the described implementations. The server cluster 200 comprises a primary server 202 having a processor 204 and a monitor 205, a first secondary server 206 having a processor 208 and a monitor 209, and a second secondary processor 210 having a processor 212 and a monitor 213. The monitors are software devices that collect server parameter values while the server cluster 200 is in operation. The server cluster 200 communicates with a master client 214 via a communications connection 216. It is noted that several clients (not shown) may be connected to the server cluster 200. However, only one client is selected by the user to be the master client 214. The master client 214 includes a simulation test program 217. The function of the master client 214 and the simulation test program 217 will be discussed in greater detail below.

The primary server 202 also includes a memory 218 and runs an operating system 220. The operating system 220 provides resource management for primary server 202 resources. The memory 218 of the primary server 202 includes a cluster controller 222, which controls communications between the primary server 202 and the secondary servers 206, 210 and between the server cluster 200 and the network 214. To accomplish this, the cluster controller 222 is provided with a communications program 224.

A capacity planner 226 is included in the cluster controller 222. The function of the capacity planner 226 and its components will be described in greater detail below. Generally, the capacity planner 226 comprises benchmark data 228 in which data collected from the server cluster 200 is stored, a calculation

1 module 230 which stores the equations necessary to derive server resource
2 utilization estimates, and plans 232 which stores recommendations that may be
3 made to improve operational configuration of the server cluster. This file of
4 recommendations is pre-defined by the manufacturer to list all the possible
5 recommendations developed for the server cluster 200. In addition, plans 232 may
6 be updated via a version upgrade or through a connection to the Internet.

7 In addition, the capacity planner 226 includes a user interface 234 and an
8 ISAPI filter 236. The user interface 234 provides areas wherein a user of the
9 server cluster 200 in general and, more specifically, the capacity planner 222 can
10 enter server parameter values and/or a specified load for which the user wants to
11 see server resource utilization and recommendations. The ISAPI filter 236 is used
12 to collect actual server parameter values from the server cluster 200 while the
13 server cluster 200 is operating. It is noted that the filter need not be an ISAPI
14 filter, but can be any type of filter capable of performing the functions listed
15 herein.

16 The capacity planner 222 includes a load simulation tool 238 which is used
17 to construct simulation scripts - such as the simulation test program 217 - that,
18 when run on the master client 214, simulates, plays or replays a server load
19 scenario using actual operating conditions recorded from the server cluster 200.
20 The use of the load simulation tool 238 is described in further detail below.

21 The implementations and functions of the components of the server cluster
22 200 outlined above will become more clear as the discussion progresses with
23 continuing reference to the components of Fig. 2.

24 The server resources that are discussed herein are: (1) processor utilization
25 (also referred to as CPU utilization), wherein the processor utilization for a given

load is expressed as a percentage of total processing power available; (2) memory utilization, expressed as a percentage of total memory available is determined by multiplying the memory required for each request by the number of requests; (3) communication bandwidth utilization, expressed as a percentage of the average throughput per bytes per second in relation to the total communication bandwidth available; and (4) general server utilization, expressed as a ratio between a current service rate (number of requests per second served) and the maximum possible service rate (maximum number of requests the server is capable of serving). The general server utilization is less specific than showing the processor, bandwidth, and memory utilization, but it is useful for viewing resource constraints that do not fall under the other categories.

Fig. 3 shows a screen shot of a user interface 300 for a capacity planning worksheet, wherein the user enters the specified load, for which the user desires to observe the effects on the system of handling such a load. The user is required to manually enter several server parameter values. These server parameter values include: number of servers in the server cluster, available communications bandwidth, server name on which a simulation will be run, client name of the client that will serve as the master test client and execute a simulation script, and the name of the script that will be used to run the simulation.

To begin, the user notifies the server cluster 200 to begin collecting data. The monitors 205, 209, 213 collect data from each server 202, 206, 210. The ISAPI filter 236 collects data for other server parameters, namely for communications-related parameters such as number of incoming requests and average response time.

1 be constructed for any of the server resource utilization estimates. As the load
2 increases to point 502 on the load axis, the utilization curve 500 reaches a point
3 504 which can be considered to be the maximum load that can be handled by the
4 server 202.

5 The user is may increase the load via the user interface 300, and re-run the
6 script using the higher load value. A situation will arise in which an increase in
7 the load will not result in an increase of the rate at which the load is handled. This
8 is the maximum load 502 which the server can handle. The load (L) at this point
9 is used in the resource utilization estimate calculations below.

10 General server utilization is derived by solving:

$$U = \frac{L}{X}$$

14 wherein:

15 U = general server utilization;

16 L = specified load; and

17 X = maximum load that can be handled by the server cluster 200.
18
19
20
21
22
23
24
25

Processor utilization is derived by solving:

$$U_{CPU} = \frac{a}{e^{b \cdot L}}$$

wherein:

U_{CPU} is processor utilization;

L is the specified load; and

a and b are processor regression constants derived from applying linear regression methodology to several load/utilization (x,y) pairs measured during the test.

Communications bandwidth utilization is derived by solving:

$$U_B = \frac{F_{TCP}}{B} \cdot (c + d \cdot L)$$

wherein:

U_B is communication bandwidth utilization;

F_{TCP} is a transmission overhead factor that, when applied to a certain size page, results in the actual bandwidth necessary to transmit the page;

L is the specified load;

B is the total communication bandwidth available; and

c and d are bandwidth regression constants derived from applying linear regression methodology to several load/utilization (x,y) pairs measured during the test.

1
2 The memory utilization is derived by first solving the following equation to
3 determine the number of concurrent connections:

4
5
$$N = \frac{L}{(X - L)} + S1 \cdot L$$

6

7 wherein:

8 *N* is the number of concurrent connections;

9 *L* is the specified load;

10 *X* is the maximum load that can be handled by the server cluster 200; and

11 *S1* is a connection memory factor that is the adjusted average of the
12 incoming connections at different speeds. For example, suppose that the ISAPI
13 filter 236 has measured the following percentages for connection types:
14

15 56K: 50%

16 ADSL: 20% ***question: what relation to screen shot? ISDN? ***

17 T1: 20%

18 T3: 10%.

19 Then *S1* is the adjusted average of these connection speeds:

20 56K: 0.5 * 5.6 = 2.8 KBytes/sec +

21 ADSL: 0.2 * 30 = 6 KBytes/sec +

22 T1: 0.2 * 150 = 30 KBytes/sec +

23 T3 0.1 * 4500 = 450 KBytes/sec = 488.8 KBytes/sec.

24 Then *S1* = 488.8 KBytes/second.
25

1 The memory utilization is thus derived by solving:

2

$$3 \quad U_M = \frac{N \cdot (M_{TCP} + M_{IISStruct}) + M_{OS} + M_{IIS}}{4 \quad M}$$

5 wherein:

6 U_M is memory utilization;

7 N is the number of concurrent connections;

8 M_{TCP} is an amount of memory for TCP buffers (32 KB per connection);

9 M_{IIS} is the amount of memory required by a server communication program
10 (50 MB for IIS);

11 $M_{IISStruct}$ is the amount of memory necessary to support communications
12 program data structures associated with each connection (50 KB per connection
13 for IIS);

14 M_{OS} is the amount of memory required by a server operating system (64
15 MB for Windows® NT by Microsoft® Corp.) and

16 M is the amount of total memory available.

17 It is noted that some figures have been used that are specific to IIS, the
18 communications program 224 used for purposes of this discussion. However, it is
19 noted that these numbers may be different for different communications programs.

20

21 **Conclusion**

22 The described implementations advantageously provide for capacity
23 planning for a server-client system and, particularly, to a server cluster within a
24 server-client system. The load simulation tool is an extremely accurate tool for
25 determining the maximum load handled by a server. The maximum load can then

